# On the Automatic Analysis of Learner Corpora

Modeling between surface features
and linguistic abstraction

Detmar Meurers
Universität Tübingen

based on joint research with
Serhiy Bykh and Julia Krivanek

CILC 2012: 4th International Conference on Corpus Linguistics
Jaén, March 22–24, 2012

## Overview

- Motivations behind analyzing learner language
- Linguistic categories for learner language
- Experimentally exploring the space between surface-based features and linguistic abstractions
  → Native language classification

## Why Analyze Learner Language?

- Second Language Acquisition (SLA) research is aimed at understanding
  - how languages are acquired
  - and how language works
- empirical basis: analysis of learner data
  - Data collected in corpora can provide empirical insights for the development & validation of linguistic theories.
- Analysis of learner language data also helps document and advance our understanding of
  - student abilities and needs
  - teaching methods and tools

  in Foreign Language Teaching and Learning (FLTL) and Intelligent Computer Assisted Language Learning (ICALL).

## Learner Data in SLA Research

An example: Clahsen & Muysken (1986)

- They studied the acquisition of German word order by native speakers of Romance languages.
- Stages of acquisition:
  1. S (Aux) V O
  2. (AdvP/PP) S (Aux) V O
  3. S V[+fin] O V[-fin]
  4. XP V[+fin] S O
  5. S V[+fin] (Adv) O
  6. *dass* S O V[+fin]

  | | | | | |
  |---|---|---|---|---|
  | Stage 2 example: | *Früher* | *ich* | *kannte* | *den Mann* |
  | | earlier$_{AdvP}$ | I$_{S}$ | knew$_{V}$ | [the man]$_{O}$ |
  | Stage 4 example: | *Früher* | *kannte* | *ich* | *den Mann* |
  | | earlier$_{AdvP}$ | knew$_{V[+fin]}$ | I$_{S}$ | [the man]$_{O}$ |

- How is the data characterized?
  - *lexical* and *syntactic categories* and *functions*
  - some acquisition stages are well-formed, others ill-formed

# Corpus Annotation for SLA Research

- SLA research essentially observes the occurrence and correlations of linguistic properties
- Corpus-based research can make use of linguistic annotation to support the identification of
  - characteristic, criterial features of language development (e.g., Hawkins & Buttery 2010)
  - quantitative measures of language development **C**omplexity, **A**ccuracy & **F**luency (Housen & Kuiken 2009)
  - overuse/underuse of linguistic material (Wiersma et al. 2011, Hirschmann et al. 2010)

⇒ What is involved in linguistically annotating learner corpora (automatically)?

# Annotation of Linguistic Properties

- Annotation schemes for native language corpora have been developed for a wide range of linguistic properties:
  - part-of-speech, morphology
  - syntactic constituency, lexical dependency structures
  - semantics (word senses, coreference), discourse structure
- An annotation scheme is only as good as the distinctions it reliably supports making based on evidence in corpus.
  - E.g., particle vs. preposition dropped in PTB tagset
  - More classes can actually be more reliable if they are more coherent in terms of their observable properties.
    - cf. BNC Tag Enhancement Project (CLAWS7 ↦ CLAWS5)
- Which linguistic categories are
  - appropriate for learner language,
  - relevant for answering research questions,
  - and can be reliably annotated?

# Appropriate categories for learner language

Parts-of-speech (Díaz Negrillo, Meurers, Valera & Wunsch 2010)

From the NOSE learner corpus (Díaz Negrillo 2009):

(1) *RED helped him **during** he was in the prison.*
- stem: preposition
- distribution: conjunction

(2) *you can find a big **vary** of beautiful beaches*
- stem: verb
- distribution: noun

(3) *one of the favourite places to visit for many **foreigns**.*
- stem: adjective
- distribution, morphology: noun

(4) *to be **choiced** for a job*
- stem: noun or adjective
- distribution, morphology: verb

# Systematic POS for Learner Language

- A single POS tag from a standard native tagset fails to systematically identify properties of learner language.
- Better: tripartite POS encoding of observable properties
  - distribution, stem, morphology

  → supports identification of mismatches in linguistic encoding
- The value of identifying such mismatches systematically is confirmed by recent SLA research (Zyzik & Azevedo 2009)
  - L2 learners are shown to have
    - difficulty distinguishing between word classes among semantically related lexical forms
    - limited ability to interpret syntactic and morphological cues

# On the nature of categories for learner language

- Annotating learner language with the standard annotation schemes developed for native language can hide important learner language characteristics.
- Comparative fallacy: “the mistake of studying the systematic character of one language by comparing it to another.” (Bley-Vroman 1983, p. 6)
- Essentially trying to analyze a “non-canonical variety” using a “robust” version of the canonical grammar.
  - divergences from norm annotated as errors
- Issue more general than language acquisition research:
  - Eurocentrism in field work (Gil 2001)
  - Variationist sociolinguistics:
    - Importance of defining variation to be studied and when exactly an instance is counted as one of the variants.

# On the nature of categories for learner language

## Between representing variation and robustness

- Where do linguistic categories come from?
  - Categories result from generalizations, which require a significant amount of comparable data to be made.
- How fine grained should they be?
  - The category system used must be sufficiently fine grained for the variation we want to identify and analyze.
  - Robustness needed to *ignore* other variation in the realization of a category to be identified.

→ To provide access to the right level of abstraction for a range of research questions: multiple levels of annotation

# On the nature of categories for learner language

## Consequences for syntactic annotation

- Idea: break down constituency in terms of
  - overall topology of a sentence (Hirschmann et al. 2007)
  - chunks (Abney 1997)
  - dependencies
    - dissociation of morphological, syntactic, and semantic dependencies (cf. also Meaning Text Theory, Mel’čuk 1988)
- Dependency analysis of learner language:
  - surface-evidence based (Dickinson & Ragheb 2009)
    - goal: fine-grained record of morphological & syntactic evid.
  - canonical dependencies (MacWhinney 2008; Rosén & Smedt 2010; Ott & Ziai 2010; Hirschmann et al. 2010)
    - goal: robustly abstract away from learner specific forms
    - e.g., in CoMiC: robust construction of semantics for comparing the meaning of answers to reading comprehension questions (Hahn & Meurers 2011)

# On developing an experimental testbed

- How can we find out more about the informativeness of the surface forms and linguistic abstractions?

  → Set up a classification experiment which allows us to quantify impact of different features.
- An interesting candidate:
  Identifying the native language (L1) of a non-native text.
- *Transfer is the influence resulting from similarities and differences between the target language and any other language that has been previously [. . . ] acquired.* (Odlin 1989, p. 27)
  - L1 Transfer occurs at many levels:
    lexical, syntactic, discourse, . . .

# Two strands of experiments

- Data-driven approach with Serhiy Bykh:
  - from surface forms to part-of-speech
- Theory-driven approach with Julia Krivanek:
  - syntactic alternations (Levin 1993) as a linguistic perspective on the data

# Data-driven approach with Serhiy Bykh

## Corpus used

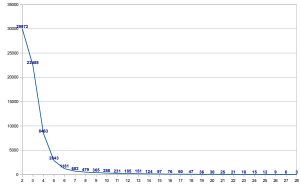
- International Corpus of Learner English (ICLE V.2, Granger et al. 2009)
  - argumentative essays written by higher intermediate to advanced learners of English, several mother tongues
- Used a subcorpus with seven native languages:
  - Bulgarian, Czech, French, Russian, Spanish, Chinese, Japanese
- 95 texts per language
  - 70 for training, 25 for testing
  - each text is between 500 and 1000 words long

⇒ For each text in the test set, determine the native language of the writer.

# Approach 1

## Setup

- efficiently identify all recurring surface forms
  - cf. variation n-gram approach to corpus annotation error detection (Dickinson & Meurers 2003, 2005; Boyd et al. 2007, 2008)
- extract all sequences of words (n-grams) which occur in at least two essays of the training corpus
  - 67.905 n-grams of length 2–28



# Approach 1

## Example features

- 2-grams: *aspect of, europeans but, would reduce, becoming the, teacher without, ago he, the team, see to, tv and, hunt and, into debts*, . . .
- 3-grams: *that smoking is, is capable of, of what they, real world the, leaves much to, of so called, their health and, to know and, need for a, difficult to accept*, . . .
- . . .
- 15-grams: *breathing secondhand smoke increase the risk of lung cancer and heart disease by about 25, dominated by science technology and industrialisation there is no longer a place for dreaming and* , . . .

# Approach 1

## Results

- We trained an SVM classifier (LIBLINEAR, Fan et al. 2008) on the 490 essays in the training set and tested on the 175 documents in the test set.
- use each recurring n-gram as a binary feature:
  - 1 if it occurs in the text, 0 if not
- Result: 87,4% accuracy of classification
  - Random baseline for seven language classes: 14.3%
  - Wong & Dras (2009): 73.7%
- What happens if we abstract away from the words within each n-gram feature
  - to words with the same part-of-speech?
  - to any words occurring there?

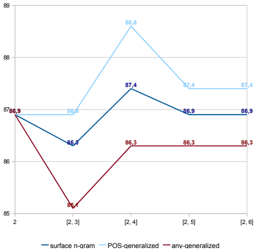
# Approach 1

## Example POS-generalized features

- 3-grams: *each JJ it, environment IN which, and DT which, family RB at, a NN this, few NNS later, attract JJR people, each JJ in, number IN crimes, imagination NN is, way CC have, on DT day*, . . .
- 4-grams: *they VBP IN the, for JJ NN to, different NNS IN view, pros CC NNS in, would VB RB longer, all IN DT we, while PRP VBP young, heart NN IN about, is DT RBS significant, in DT NN market*, . . .

# Approach 1

## Overall results



- Generalization to POS classes improves result, whereas non-linguistic generalization does not.
- Success, but it is hard to qualitatively interpret features in terms of L1 transfer!

# An alternative

- Word-based surface features always encode form and meaning together.
  - requires very high number of features to be applicable to unseen data, across domains/topics
- Can we abstract away from the meaning to be expressed to choices in the linguistic system?
  - Idea: Study where the linguistic system provides multiple ways to express the same meaning.
    - similar to variationist sociolinguistics (though typically based on pronunciation variation, lexical choice there)
- How about valence alternations (Levin 1993)?
  - e.g., Dative Alternation

    (5) a. *He gave the book to John.*

    b. *He gave John the book.*

  - Popular topic in linguistics, but so far little corpus-based SLA work (but cf. Callies & Zaytseva 2011).

# Theory-driven approach with Julia Krivanek
## Setup

- Corpus used:
  - L1 Chinese from ICLE (V.2, Granger et al. 2009)
  - native English essays from LOCNESS (http://www.uclouvain.be/en-cecl-locness.html)
- Goal: binary classification into non-native vs. native
  - training: 600 documents, evenly split
  - testing: 120 documents, evenly split
- We focused on 21 alternation which can be reliably identified given syntactic annotation.
  - about 1/5 of the ones given in Levin (1993)

# Theory-driven approach
## Identifying alternations

- Easy to identify: *as*-Alternation

(6) a. *He appointed him press secretary.*
appoint + NP + NP
b. *He appointed him as press secretary.*
appoint + NP + PP(as)

- More difficult to identify: Simple Reciprocal Alternation

(7) a. *Anna agreed with John*
b. *Anna and John agreed.*

(8) *# Anna agreed with the argument.*

→ additional information (e.g., animacy) relevant

# Theory-driven approach
## Results

- syntactically annotated corpus with Bitpar (Schmid 2004)
  - trained on enriched WSJ from PennTreebank
  - → lexical categories contain subcategorization information
- identify syntactic alternations using tgrep2 patterns
- 21 binary alternations: 42 features per document
- features: choices per class made in a document
  - for each class, record relative frequency of choices
  - e.g., for document with 3 instances of a class: $\frac{2}{3}$, $\frac{1}{3}$
- 63.33% Accuracy (SVM: Weka SMO)
  - average document length only 790 words
  - not enough instances of relevant patterns per document!
  - when pooling 5 documents (120 train, 24 test): 70.83%

# Theory-driven approach
## . . . with a data-driven twist

- for each verb, record its selection patterns in the corpus
  - define classes consisting of all verbs with the same set of syntactic realization alternatives
- Corpora: L1 English (LOCNESS), L1 Chinese (ICLEv2)
  - training: 600 documents, evenly split
  - testing: 120 documents, evenly split
- Result: 72.5% accuracy (SVM: Weka SMO)
  - 87.5% when pooling 5 documents (120 train, 24 test)
  - 95.83% with alternative definition of verb classes

# Qualitative analysis of underuse/overuse

## Patterns

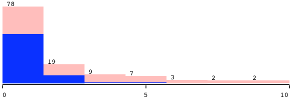
- overused in learner language: *provide* NP NP

  (9) *Universities provides us a chance to live.* (ICLEv2)

- underused in learner language: *see* NP *as* NP

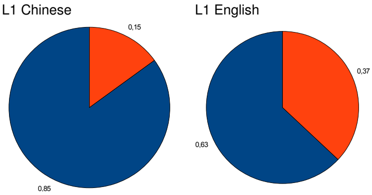
  (10) *Now we see it as being absurd in America that women did not have a right to vote.* (LOCNESS)

- general “V NP as NP” also underused by learners (blue)



# Qualitative analysis: distinctive alternations

## Locative Preposition Drop Alternation

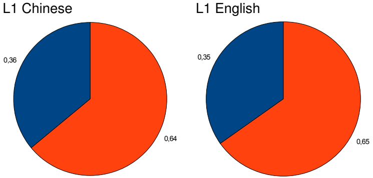


🟥 V-PPloc (Martha climbed up the mountain.)

🟦 V-NP (Martha climbed the mountain.)

# Qualitative analysis: indistinctive alternations

## Dative Alternation



🟦 V-NP-NP

🟥 V-NP-to-NP

# Conclusion (I)

- We started with sketching the role of learner language corpora in SLA and FLTL.
- Linguistic annotation is motivated by the need to support effective querying for relevant patterns.
  - Corpus annotation provides access to classes of data,
  - but annotated classes need to be appropriate for the type of language and research question at hand.
- The issue is particularly difficult for the individual interlanguage systems in language development.
  - standard annotation schemes can hide characteristics
  - need to balance robustness vs. variation to be captured
  - multilayer annotation useful to support range of research questions

# Conclusion (II)

- L1 classification as experimental sandbox for exploring impact of features between surface & linguistic abstraction.
  - Approach 1 with Serhiy Bykh:
    - data-driven: surface n-gram based
    - but: value of part-of-speech generalization
  - Approach 2 with Julia Krivanek:
    - theory-driven: alternation-based
    - but: value of data-driven class definitions
- General research direction:
  - Where can linguistic abstractions be shown to matter?
  - When does it pay off to identify a class or a rule instead of just storing all experience?

# References

Aarts, J. & S. Granger (1998). Tag Sequences in Learner Corpora: a Key to Interlanguage Grammar and Discourse. In S. Granger (ed.), *Learner English on Computer*, London; New York: Longman, pp. 132–141.

Abney, S. (1997). Partial Parsing via Finite-State Cascades. *Natural Language Engineering* 2, 337–344. URL http://www.vinartus.net/spa/97a.pdf.

Amaral, L., V. Metcalf & D. Meurers (2006). Language Awareness through Re-use of NLP Technology. Pre-conference Workshop on NLP in CALL – Computational and Linguistic Challenges. CALICO 2006. May 17, 2006. University of Hawaii. URL http://purl.org/net/icall/handouts/calico06-amaral-metcalf-meurers.pdf.

Amaral, L. & D. Meurers (2008). From Recording Linguistic Competence to Supporting Inferences about Language Acquisition in Context: Extending the Conceptualization of Student Models for Intelligent Computer-Assisted Language Learning. *Computer-Assisted Language Learning* 21(4), 323–338. URL http://purl.org/dm/papers/amaral-meurers-call08.html.

Amaral, L. & D. Meurers (2009). Little Things With Big Effects: On the Identification and Interpretation of Tokens for Error Diagnosis in ICALL. *CALICO Journal* 26(3), 580–591. URL http://purl.org/dm/papers/amaral-meurers-09.html.

Amaral, L. & D. Meurers (2011). On Using Intelligent Computer-Assisted Language Learning in Real-Life Foreign Language Teaching and Learning. *ReCALL* 23(1), 4–24. URL http://purl.org/dm/papers/amaral-meurers-10.html.

Amaral, L., D. Meurers & R. Ziai (2011). Analyzing Learner Language: Towards A Flexible NLP Architecture for Intelligent Language Tutors. *Computer-Assisted Language Learning* 24(1), 1–16. URL http://purl.org/dm/papers/amaral-meurers-ziai-10.html.

Bailey, S. & D. Meurers (2008). Diagnosing meaning errors in short answers to reading comprehension questions. In J. Tetreault, J. Burstein & R. D. Felice (eds.), *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications (BEA-3) at ACL'08*. Columbus, Ohio, pp. 107–115. URL http://aclweb.org/anthology/W08-0913.

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning* 33(1), 1–17. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-1770.1983.tb00983.x/pdf.

Boyd, A., M. Dickinson & D. Meurers (2007). Increasing the Recall of Corpus Annotation Error Detection. In *Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories (TLT-07)*. Bergen, Norway. URL http://purl.org/dm/papers/boyd-et-al-07b.html.

Boyd, A., M. Dickinson & D. Meurers (2008). On Detecting Errors in Dependency Treebanks. *Research on Language and Computation* 6(2), 113–137. URL http://purl.org/dm/papers/boyd-et-al-08.html.

Boyd, A. & D. Meurers (2008). On Diagnosing Word Order Errors. Poster presented at the CALICO Pre-Conference Workshop on Automatic Analysis of Learner Language. URL http://purl.org/net/calico-workshop-abstracts.html#6.

Callies, M. & E. Zaytseva (2011). The Corpus of Academic Learner English (CALE): A new resource for the study of lexico-grammatical variation in advanced learner varieties. In *Hedeland, Hanna and Thomas Schmidt and Kai Wörner*, Multilingual Resources and Multilingual Applications (Hamburg Working Papers in Multilingualism B 96), pp. 51–56.

Clahsen, H. & P. Muysken (1986). The availability of universal grammar to adult and child learners - a study of the acquisition of German word order. *Second Language Research* 2(2), 93–119. URL http://slr.sagepub.com/content/2/2/93.abstract.

Covington, M. A., C. He, C. Brown, L. Naçi & J. Brown (2006). *How complex is that sentence? A proposed revision of the Rosenberg and Abbeduto D-Level Scale*. Computer Analysis of Speech for Psychological Research (CASPR) Research Report 2006-01, The University of Georgia, Artificial Intelligence Center, Athens, GA. URL http://www.ai.uga.edu/caspr/2006-01-Covington.pdf.

Dickinson, M. & W. D. Meurers (2003). Detecting Errors in Part-of-Speech Annotation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*. Budapest, Hungary, pp. 107–114. URL http://purl.org/dm/papers/dickinson-meurers-03.html.

Dickinson, M. & W. D. Meurers (2005). Detecting Errors in Discontinuous Structural Annotation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. pp. 322–329. URL http://aclweb.org/anthology/P05-1040.

Dickinson, M. & M. Ragheb (2009). Dependency Annotation for Learner Corpora. In *Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories (TLT-8)*. Milan, Italy. URL http://jones.ling.indiana.edu/~mdickinson/papers/dickinson-ragheb09.html.

Díaz Negrillo, A. (2009). *EARS: A User's Manual*. Munich, Germany: LINCOM Academic Reference Books.

Díaz Negrillo, A., D. Meurers, S. Valera & H. Wunsch (2010). Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum* 36(1–2), 139–154. URL http://purl.org/dm/papers/diaz-negrillo-et-al-09.html. Special Issue on Corpus Linguistics for Teaching and Learning. In Honour of John Sinclair.

Elghafari, A., D. Meurers & H. Wunsch (2010). Exploring the Data-Driven Prediction of Prepositions in English. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*. Beijing, China, pp. 267–275. URL http://aclweb.org/anthology/C10-2031.

Fan, R., K. Chang, C. Hsieh, X. Wang & C. Lin (2008). LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research* 9, 1871–1874.

Gil, D. (2001). Escaping Eurocentrism: Fieldwork as a Process of Unlearning. In P. Newman & M. Ratliff (eds.), *Linguistic Fieldwork*, Cambridge University Press, pp. 102–132.

Granger, S., E. Dagneaux, F. Meunier & M. Paquot (2009). *International Corpus of Learner English Version 2*. Presses Universitaires de Louvain.

Hahn, M. & D. Meurers (2011). On deriving semantic representations from dependencies: A practical approach for evaluating meaning in learner corpora. In *Proceedings of the Intern. Conference on Dependency Linguistics (DEPLING 2011)*. Barcelona. URL http://purl.org/dm/papers/hahn-meurers-11.html.

Hawkins, J. A. & P. Buttery (2009). Using Learner Language from Corpora to Profile Levels of Proficiency – Insights from the English Profile Programme. In *Studies in Language Testing: The Social and Educational Impact of Language Assessment*, Cambridge: Cambridge University Press.

Hawkins, J. A. & P. Buttery (2010). Criterial Features in Learner Corpora: Theory and Illustrations. *English Profile Journal* .

Hirschmann, H., S. Doolittle & A. Lüdeling (2007). Syntactic annotation of non-canonical linguistic structures. In *Proceedings of Corpus Linguistics 2007*. Birmingham. URL http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/neu2/mitarbeiter-innen/anke/pdf/HirschmannDoolittleLuedelingCL2007.pdf.

Hirschmann, H., A. Lüdeling, I. Rehbein, M. Reznicek & A. Zeldes (2010). Syntactic Overuse and Underuse: A Study of a Parsed Learner Corpus and its Target Hypothesis. Presentation given at the Treebanks and Linguistic Theory Workshop.

Housen, A. & F. Kuiken (2009). Complexity, Accuracy, and Fluency in Second Language Acquisition. *Applied Linguistics* 30(4), 461–473. URL http://applij.oxfordjournals.org/content/30/4/461.full.pdf.

Krivanek, J. & D. Meurers (2011). Comparing Rule-Based and Data-Driven Dependency Parsing of Learner Language. In *Proceedings of the Intern. Conference on Dependency Linguistics (DEPLING 2011)*. Barcelona.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*, vol. 348. Chicago, IL: University of Chicago Press.

Lu, X. (2008). Automatic measurement of syntactic complexity using the revised developmental scale. In *Proceedings of the 21st International Florida Artificial Intelligence Research Society Conference (FLAIRS-08)*. Coconut Grove, FL: AAAI Press.

Lu, X. (2009). Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics* 14(1), 3–28.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15(4), 474–496.

MacWhinney, B. (2008). Enriching CHILDES for morphosyntactic analysis. In H. Behrens (ed.), *Corpora in Language Acquisition Research: History, Methods, Perspectives*, Amsterdam and Philadelphia: John Benjamins, vol. 6 of *Trends in Language Acquisition Research*, pp. 165–197. URL http://childes.psy.cmu.edu/grasp/morphosyntax.doc.

Mel'čuk, I. (1988). *Dependency Syntax: Theory and Practice*. State University of New York Press. URL http://books.google.com/books?id=diq29vrjAa4C&lpg=PR13&ots=ZcCJBmEA7g&dq=Dependency20Syntax3A20Theory20and20Practice&lr&pg=PR13#v=onepage&q&f=false.

Metcalf, V. & D. Meurers (2006a). Generating Web-based English Preposition Exercises from Real-World Texts. URL http://purl.org/net/icall/handouts/eurocall06-metcalf-meurers.pdf. EUROCALL 2006. Granada, Spain. September 4–7, 2006.

Metcalf, V. & D. Meurers (2006b). When to Use Deep Processing and When Not To – The Example of Word Order Errors. URL http://purl.org/net/icall/handouts/calico06-metcalf-meurers.pdf. Pre-conference Workshop on NLP in CALL – Computational and Linguistic Challenges. CALICO 2006. May 17, 2006. University of Hawaii.

Meurers, D. (2009). On the Automatic Analysis of Learner Language: Introduction to the Special Issue. *CALICO Journal* 26(3), 469–473. URL http://purl.org/dm/papers/meurers-09.html.

Meurers, D., N. Ott & R. Ziai (2010a). Compiling a Task-Based Corpus for the Analysis of Learner Language in Context. In *Pre-Proceedings of Linguistic Evidence*. Tübingen, pp. 214–217. URL http://purl.org/dm/papers/meurers-ott-ziai-10.html.

Meurers, D., R. Ziai, L. Amaral, A. Boyd, A. Dimitrov, V. Metcalf & N. Ott (2010b). Enhancing Authentic Web Pages for Language Learners. In *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-5) at NAACL-HLT 2010*. Los Angeles: Association for Computational Linguistics. URL http://purl.org/dm/papers/meurers-ziai-et-al-10.html.

Meurers, W. D. (2005). On the use of electronic corpora for theoretical linguistics. Case studies from the syntax of German. *Lingua* 115(11), 1619–1639. URL http://purl.org/dm/papers/meurers-03.html.

Meurers, W. D. & S. Müller (2009). Corpora and Syntax (Article 42). In A. Lüdeling & M. Kytö (eds.), *Corpus linguistics*, Berlin: Mouton de Gruyter, vol. 2 of *Handbooks of Linguistics and Communication Science*, pp. 920–933. URL http://purl.org/dm/papers/meurers-mueller-09.html.

Odlin, T. (1989). *Language Transfer: Cross-linguistic influence in language learning*. New York: Cambridge University Press.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics* 24(4), 492–518.

Ortega, L. & H. Byrnes (eds.) (2008). *The longitudinal study of advanced L2 capacities*. Routledge.

Ott, N. & D. Meurers (2010). Information Retrieval for Education: Making Search Engines Language Aware. *Themes in Science and Technology Education. Special issue on computer-aided language analysis, teaching and learning: Approaches, perspectives and applications* 3(1–2), 9–30. URL http://purl.org/dm/papers/ott-meurers-10.html.

Ott, N. & R. Ziai (2010). Evaluating Dependency Parsing Performance on German Learner Language. In M. Dickinson, K. Müürisep & M. Passarotti (eds.), *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories*. vol. 9 of *NEALT Proceeding Series*, pp. 175–186. URL http://hdl.handle.net/10062/15960.

Rosenberg, S. & L. Abbeduto (1987). Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults. *Applied Psycholinguistics* 8, 19–32. URL http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=2607480&fulltextType=RA&fileId=S0142716400000047.

Rosén, V. & K. D. Smedt (2010). Syntactic Annotation of Learner Corpora. In H. Johansen, A. Golden, J. E. Hagen & A.-K. Helland (eds.), *Systematisk, variert, men ikke tilfeldig. Antologi om norsk som andrespråk i anledning Kari Tenfjords 60-årsdag [Systematic, varied, but not arbitrary. Anthology about Norwegian as a second language on the occasion of Kari Tenfjord's 60th birthday]*, Oslo: Novus forlag, pp. 120–132.

Scarborough, H. S. (1990). Index of Productive Syntax. *Applied Psycholinguistics* 11(1), 1–22. URL http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=2747016.

Schmid, H. (2004). Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*. Geneva, Switzerland. URL http://www.ims.uni-stuttgart.de/www/projekte/gramotron/PAPERS/COLING04/BitPar.pdf.

Wiersma, W., J. Nerbonne & T. Lauttamus (2011). Automatically Extracting Typical Syntactic Differences from Corpora. *Literary and Linguistic Computing* 26(1), 107–124. URL http://llc.oxfordjournals.org/content/26/1/107.abstract.

Wolfe-Quintero, K., S. Inagaki & H.-Y. Kim (1998). *Second Language Development in Writing: Measures of Fluency, Accuracy & Complexity*. Honolulu: Second Language Teaching & Curriculum Center, University of Hawaii at Manoa.

Wong, S. & M. Dras (2009). Contrastive analysis and native language identification. In *Australasian Language Technology Association Workshop 2009*. p. 53.

Zyzik, E. & C. Azevedo (2009). Word Class Distinctions in Second Language Acquisition. *SSLA* 31(31), 1–29. URL http://journals.cambridge.org/production/action/cjoGetFulltext?fulltextid=3981776.